

Subspace Cluster and Outlier Detection in Big Data

Klemens Böhm

Institute for Program Structures and Data Organization (IPD)

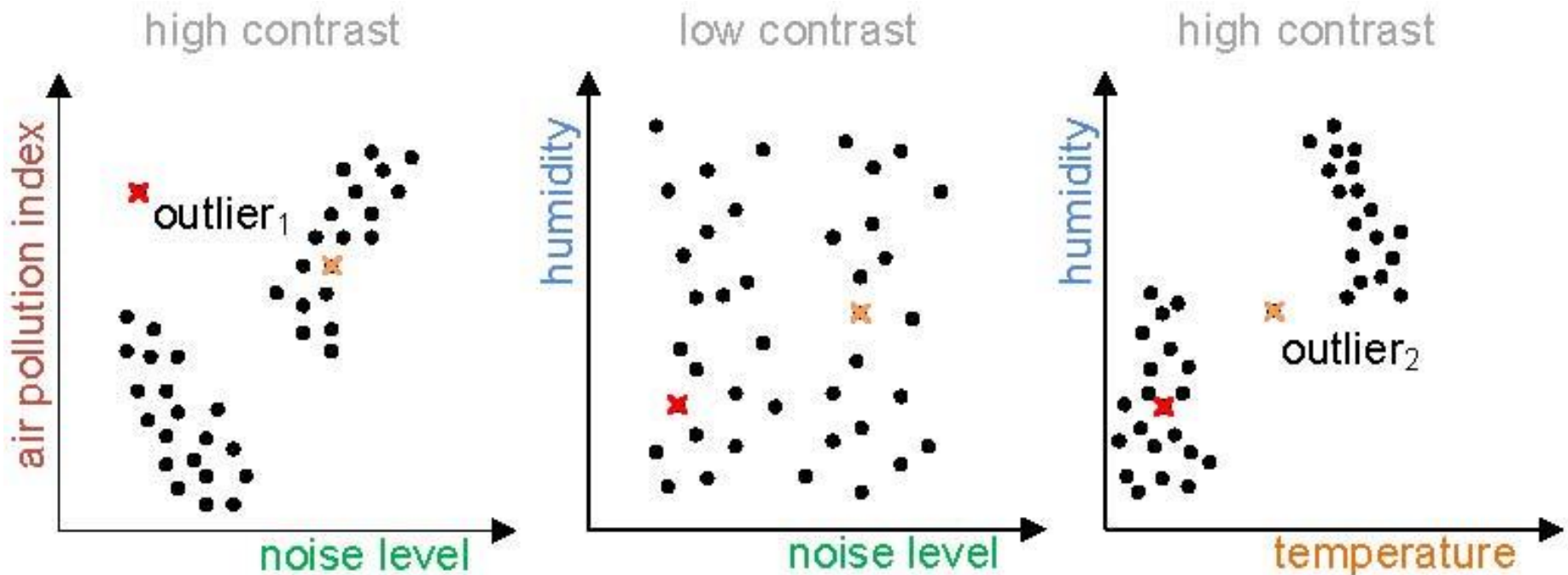
Bundesliga-Database

– Many Other Attributes Are Conceivable.

- Goals in overtime
- Goals in last 15 minutes of match
- Header goals
- Share of duels won
- Share of duels won in own penalty area
- Share of duels won in opponent's penalty area
- Share of passes which have arrived
- Share of passes over 10 meters which have arrived
- Share of passes which have arrived in opponent's half
- ...

Illustration

- Relation – attributes:
air pollution index, noise level, humidity, temperature, ...



- Domain-independence
– all this happens with data from **your** scientific discipline as well!

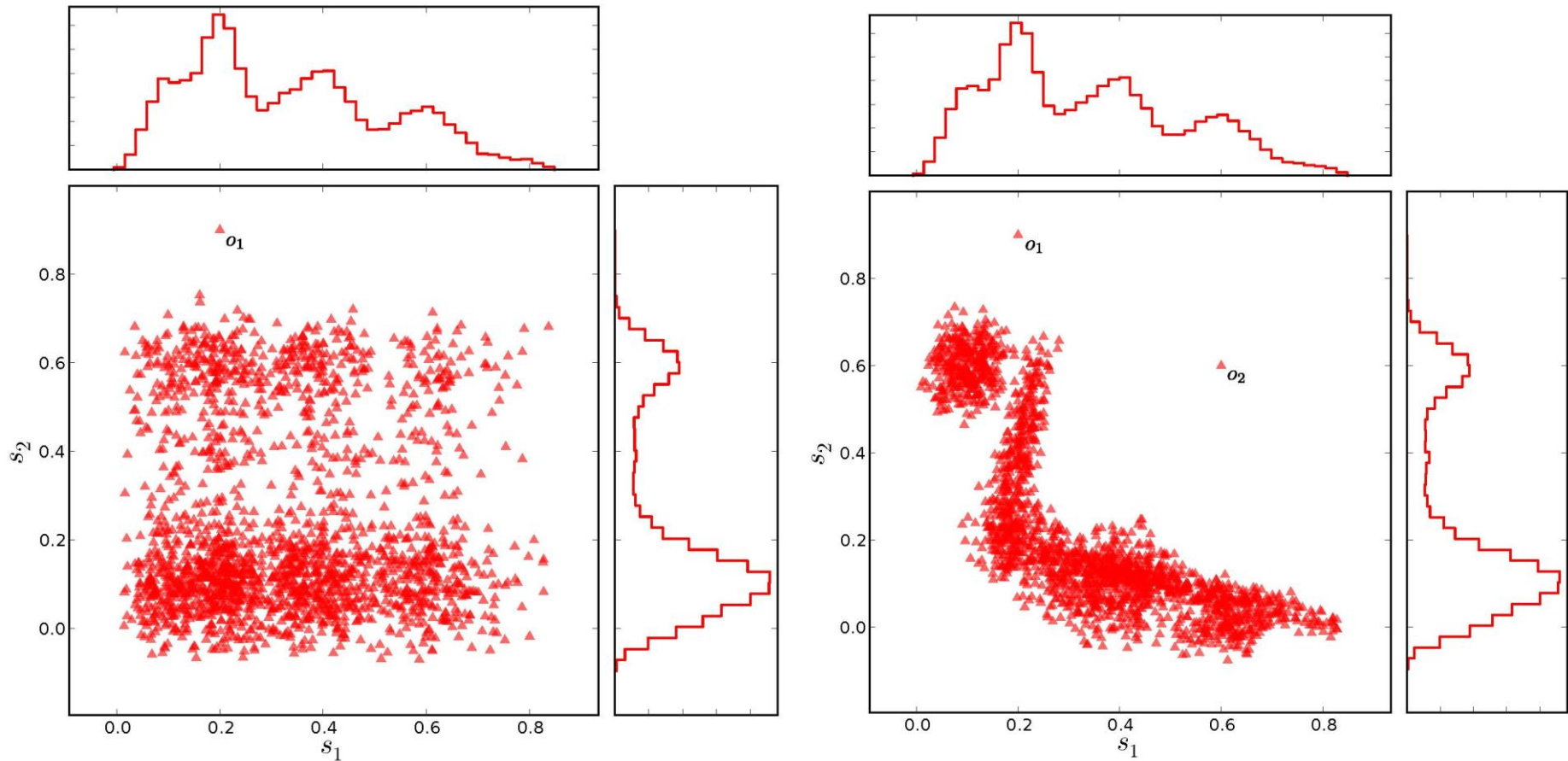
Problem Statement

- Identify subspaces that are likely to contain outliers.
 - ⇒ Quality measure for subspaces sought.
I.e., how promising is it to search the subspace for outliers?
 - ⇒ How to find such relevant subspaces in the first place?

- These subspaces can then be searched for outliers subsequently using conventional methods.

- Different outlier models
(and respective techniques for outlier detection) exist,
which typically yield different results.

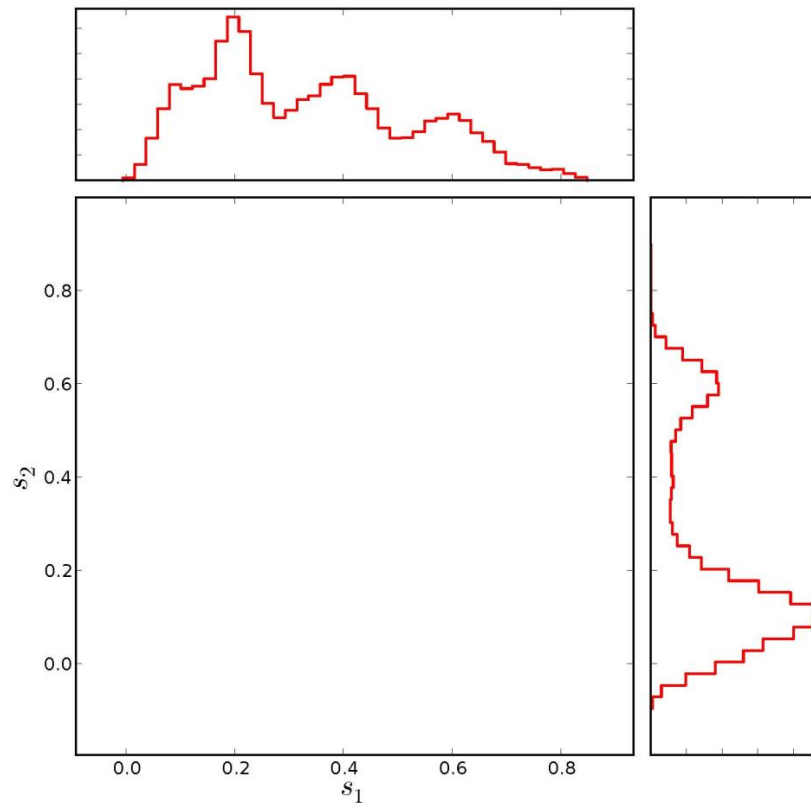
HiCS: Principle (1)



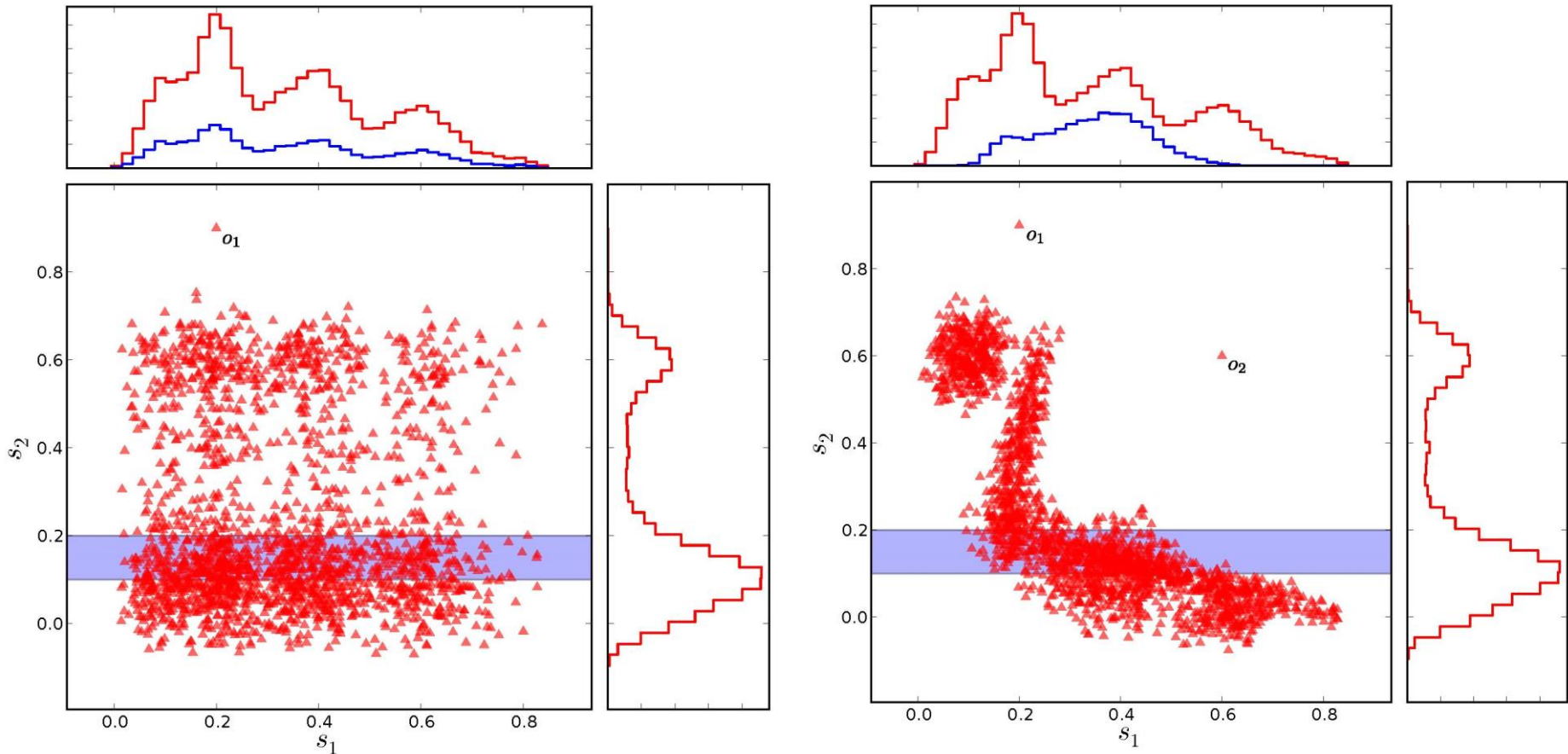
- Characteristics of data not obvious in lower-dimensional projections.
- 1D outlier vs. real 2D outlier.

HiCS: Principle (2)

- ‚unexpected behavior‘ \equiv contrast \equiv potential outliers
- Idea: search for violations of statistical independence.



HiCS: Principle (3)



■ Compare marginal distribution to conditional distribution.

How to Compute the Deviation?

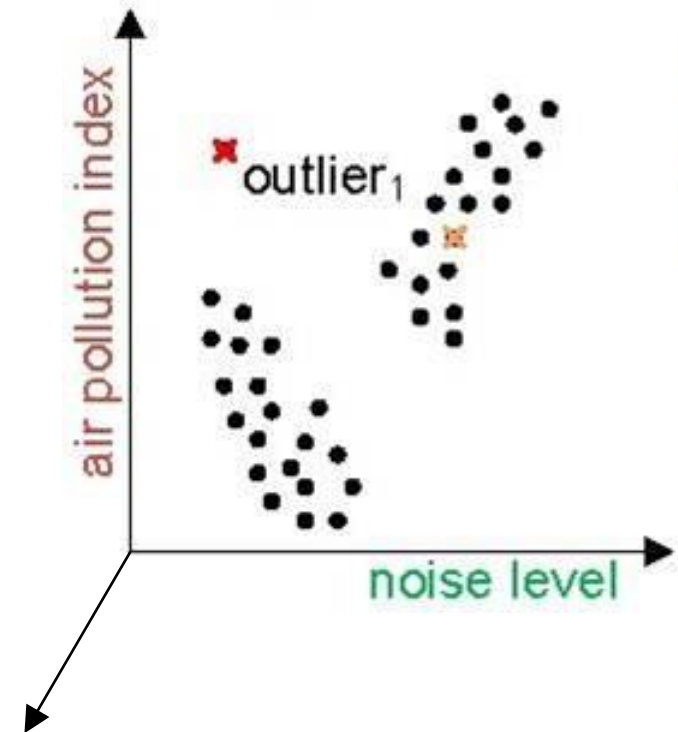
- Use established methods to compare samples (statistical test).
- Null Hypothesis H_0 :
„The samples are based on the same distribution.“
- Possible instantiations:
 - Welch-t-Test
 - Kolmogorov-Smirnov

RefOut – Motivation

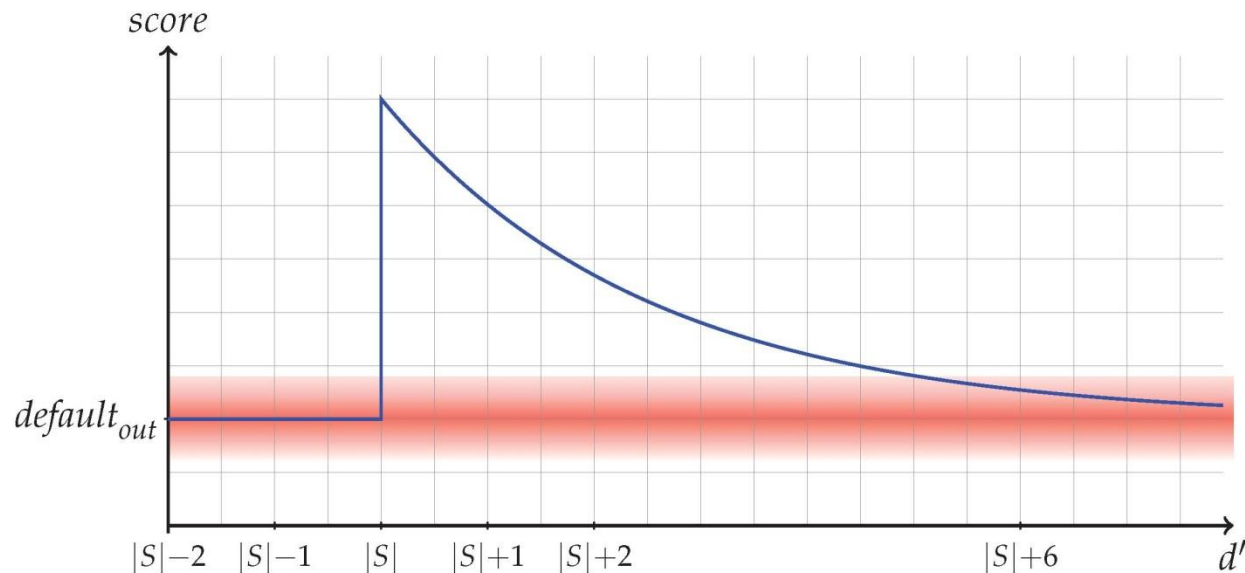
- HiCS is based on generic characteristics (data distributions or differences between distributions).
- Different outlier models exist, e.g., distance-based, density-based, angle-based.
- **According to the different outlier models, different objects – in different subspaces – are outliers.**
- Experiments of ours confirm this (very much).
- In what follows, subspace search method that takes outlier model into account.

Fundamentals

- Outlier model: $\text{score}(\bar{x}_S) \in \mathbb{R} \quad \forall \bar{x} \in \text{DB}$
- Illustration:
 - $\text{score}(x_{\{\text{noise level, air pollution index}\}})$ is large.
 - $\text{score}(x_{\{\text{noise level}\}})$ is small.
 - $\text{score}(x_{\{\text{noise level, air pollution index, d3}\}})$ is large, but smaller than $\text{score}(x_{\{\text{noise level, air pollution index}\}})$.
 - $\text{score}(x_{\{\text{noise level, air pollution index, d3, \dots, d99}\}})$ is again small.
- Normalization
 - mean value = default_{out}
 - variance (for each subspace S) = 1



Characteristics of Outliers in Subspaces



- This illustration features an idealization.
- S – subspace where current object is prominent outlier.
- $T \supset S$: Scores decrease monotonically. „dimensional curse“.

Subspace Pool

- Object under consideration in what follows:
 \bar{x} , true subspace outlier in S .
E. g., x is true subspace outlier in $S = \{\text{air pollution index, noise level}\}$.
- Given \bar{x} , how do we find S , subspace where the score of \bar{x} is maximal?
- Subspace pool $P = \text{random subset of } 2^A$.

Subspace Pool – Illustration (1)

Rank	Occurrence of attributes 1-12	Outlierness
1	Green Green Green Red Green Green Green Red Red Green	Long blue bar
2	Green Green Green Red Red Green Green Green Green	Medium-long blue bar
3	Green Green Green Green Green Red Green Red Green Red Green	Medium blue bar
4	Green Green Green Green Green Green Green Red Red Red Green Green	Medium-long blue bar
5	Red Green Green Green Green Green Green Red Green Red Green	Medium blue bar
6	Red Red Green Green Green Green Green Red Green Green Green	Medium blue bar
7	Green Green Green Red Red Green Green Red Green Green Green	Medium blue bar
8	Red Green Green Green Green Red Green Green Green Red Green	Medium blue bar
9	Red Green Red Green Green Green Green Green Red Green Green	Medium blue bar
10	Red Green Green Green Green Red Green Green Green Green Red	Medium blue bar
11	Green Green Green Red Green Green Green Red Green Green Red	Medium blue bar
12	Red Green Green Green Green Green Green Red Green Red Green	Medium blue bar
13	Green Green Green Red Green Green Red Green Green Red Green	Medium blue bar
14	Green Green Red Green Green Green Red Green Red Green Green	Medium blue bar
15	Red Green Red Green Green Green Green Red Green Green Green	Medium blue bar
16	Green Red Green Green Red Green Green Red Green Green Green	Medium blue bar
17	Green Green Red Green Green Green Green Red Green Red Green	Medium blue bar
18	Green Red Green Green Green Red Green Green Green Green Red	Medium blue bar
19	Green Green Green Red Green Green Red Red Green Green Green	Short blue bar
20	Green Green Green Red Green Red Red Green Green Green Green	Very short blue bar



Subspace Pool – Explanation of Illustration

- Dimensionality of database: $D = 12$
- Size of subspace pool: $|P| = 20$
Subspaces have dimensionality $|T| = 9$
- Dimensions of current subspace depicted in green, e.g., $T_1 = \{1, 2, 3, 4, 6, 7, 8, 9, 12\}$.

Score Discrepancy Problem

- Object under consideration in what follows:
 \bar{x} , true subspace outlier in S .
E. g., x is true subspace outlier in $S = \{\text{air pollution index, noise level}\}$.
- Subspace pool $P =$ random subset of 2^A .
- $P_S^+ = \{T \mid T \supset S \wedge T \in P\}$
- $P_S^- = \{T \mid S \not\subset T \wedge T \in P\}$

Subspace Pool – Illustration (2)

 $P_{\{1,2,3,4\}}^+$
 $P_{\{1,2,3,4\}}^-$

Rank	Occurrence of attributes 1-12	Outlierness
1	12 green, 4 red	High
2	12 green, 4 red	High
3	12 green, 4 red	High
4	12 green, 4 red	High
5	12 green, 4 red	Medium
6	12 green, 4 red	Medium
7	12 green, 4 red	High
8	12 green, 4 red	Medium
9	12 green, 4 red	Medium
10	12 green, 4 red	Medium
11	12 green, 4 red	Medium
12	12 green, 4 red	Medium
13	12 green, 4 red	Medium
14	12 green, 4 red	Medium
15	12 green, 4 red	Medium
16	12 green, 4 red	Medium
17	12 green, 4 red	Medium
18	12 green, 4 red	Medium
19	12 green, 4 red	Low
20	12 green, 4 red	Low

Subspace Pool – Illustration (3)

■ $P_{\{9,10,11,12\}}^+$



■ $P_{\{9,10,11,12\}}^-$

Rank	Occurrence of attributes 1-12	Outlierness
1	Green, Green, Green, Red, Green, Green, Green, Red, Red, Green, Green, Green	High
2	Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown	Medium
3	Green, Green, Green, Green, Green, Red, Green, Red, Green, Red, Green, Green	High
4	Green, Green, Green, Green, Green, Green, Red, Red, Red, Green, Green, Green	Medium
5	Red, Green, Green, Green, Green, Green, Green, Red, Green, Red, Green, Green	Medium
6	Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown	Medium
7	Green, Green, Green, Red, Red, Green, Red, Green, Green, Green, Green, Green	High
8	Red, Green, Green, Green, Red, Green, Green, Green, Green, Red, Green, Green	High
9	Red, Red, Green, Green, Green, Green, Green, Red, Green, Green, Green, Green	High
10	Red, Green, Green, Green, Green, Red, Green, Green, Green, Green, Red, Green	Medium
11	Green, Green, Green, Red, Green, Green, Green, Red, Green, Green, Red, Green	Medium
12	Red, Green, Green, Green, Green, Green, Green, Red, Green, Red, Green, Green	Medium
13	Green, Green, Green, Red, Green, Green, Red, Green, Green, Green, Red, Green	Medium
14	Green, Green, Red, Green, Green, Red, Green, Red, Green, Green, Green, Green	Medium
15	Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown	Medium
16	Green, Red, Green, Green, Red, Green, Green, Red, Green, Green, Green, Green	Medium
17	Green, Green, Red, Green, Green, Green, Green, Red, Green, Red, Green, Green	Medium
18	Green, Red, Green, Green, Red, Green, Green, Green, Green, Red, Green, Green	Medium
19	Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown	Low
20	Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown	Low

Score Discrepancy Problem



- Object under consideration here:
 \bar{x} , true subspace outlier in S .
 E. g., x is true subspace outlier in $S = \{\text{air pollution index, noise level}\}$.
- Subspace pool $P = \text{random subset of } 2^A$.
- $P_S^+ = \{T \mid T \supset S \wedge T \in P\}$
- $P_S^- = \{T \mid S \not\subset T \wedge T \in P\}$
- Outlier scores corresponding to P : $O = \{\text{score}(\bar{x}_T) \mid T \in P\}$
- $O_S^+ = \{\text{score}(\bar{x}_T) \mid T \in P_S^+\}$
- $O_S^- = \{\text{score}(\bar{x}_T) \mid T \in P_S^-\}$

Subspace Pool – Illustration (4)

 $O_{\{1,2,3,4\}}^+$
 $O_{\{1,2,3,4\}}^-$

Rank	Occurrence of attributes 1-12	Outlierness
1	12 green, 4 red	High (brown bar)
2	12 green, 4 red	High (brown bar)
3	12 green, 4 red	High (brown bar)
4	12 green, 4 red	High (brown bar)
5	12 green, 4 red	Low (blue bar)
6	12 green, 4 red	Low (blue bar)
7	12 green, 4 red	High (brown bar)
8	12 green, 4 red	Low (blue bar)
9	12 green, 4 red	Low (blue bar)
10	12 green, 4 red	Low (blue bar)
11	12 green, 4 red	Low (blue bar)
12	12 green, 4 red	Low (blue bar)
13	12 green, 4 red	Low (blue bar)
14	12 green, 4 red	Low (blue bar)
15	12 green, 4 red	Low (blue bar)
16	12 green, 4 red	Low (blue bar)
17	12 green, 4 red	Low (blue bar)
18	12 green, 4 red	Low (blue bar)
19	12 green, 4 red	Low (blue bar)
20	12 green, 4 red	Low (blue bar)

Subspace Pool – Illustration (5)

 $O_{\{9,10,11,12\}}^+$
 $O_{\{9,10,11,12\}}^-$

Rank	Occurrence of attributes 1-12	Outlierness
1	Green, Green, Green, Red, Green, Green, Green, Red, Red, Green, Green, Green	Blue bar (high)
2	Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown	Brown bar (low)
3	Green, Green, Green, Green, Green, Red, Green, Red, Green, Red, Green, Green	Blue bar (high)
4	Green, Green, Green, Green, Green, Green, Red, Red, Red, Green, Green, Green	Blue bar (high)
5	Red, Green, Green, Green, Green, Green, Green, Green, Red, Green, Red, Green	Blue bar (medium)
6	Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown	Brown bar (low)
7	Green, Green, Green, Red, Red, Green, Red, Green, Green, Green, Green, Green	Blue bar (high)
8	Red, Green, Green, Green, Red, Green, Green, Green, Green, Red, Green, Green	Blue bar (high)
9	Red, Red, Green, Green, Green, Green, Green, Green, Red, Green, Green, Green	Blue bar (high)
10	Red, Green, Green, Green, Green, Red, Green, Green, Green, Green, Red, Green	Blue bar (medium)
11	Green, Green, Green, Red, Green, Green, Green, Red, Green, Green, Red, Green	Blue bar (medium)
12	Red, Green, Green, Green, Green, Green, Green, Red, Green, Red, Green, Green	Blue bar (medium)
13	Green, Green, Green, Red, Green, Green, Red, Green, Green, Green, Red, Green	Blue bar (medium)
14	Green, Green, Red, Green, Green, Red, Green, Red, Green, Green, Green, Green	Blue bar (medium)
15	Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown	Brown bar (low)
16	Green, Red, Green, Green, Red, Green, Green, Red, Green, Green, Green, Green	Blue bar (medium)
17	Green, Green, Red, Green, Green, Green, Green, Red, Green, Red, Green, Green	Blue bar (medium)
18	Green, Red, Green, Green, Green, Red, Green, Green, Green, Green, Red, Green	Blue bar (medium)
19	Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown	Brown bar (low)
20	Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown, Brown	Brown bar (low)

Score Discrepancy Problem

- Object under consideration here:
 \bar{x} , true subspace outlier in S .
 E. g., x is true subspace outlier in $S = \{\text{air pollution index, noise level}\}$.
- Subspace pool $P = \text{random subset of } 2^A$.
- $P_S^+ = \{T \mid T \supset S \wedge T \in P\}$
- $P_S^- = \{T \mid S \not\subset T \wedge T \in P\}$
- Outlier scores corresponding to P : $O = \{\text{score}(\bar{x}_T) \mid T \in P\}$
- $O_S^+ = \{\text{score}(\bar{x}_T) \mid T \in P_S^+\}$
- $O_S^- = \{\text{score}(\bar{x}_T) \mid T \in P_S^-\}$

- $E[O_S^+] > E[O_S^-]$

Score Discrepancy – Problem Statement

- S however is not known. We rather want to find S .

- $\operatorname{argmax}_{S'} (E[O_{S'}^+] - E[O_{S'}^-])$

(Dimensionality of S' is given.)

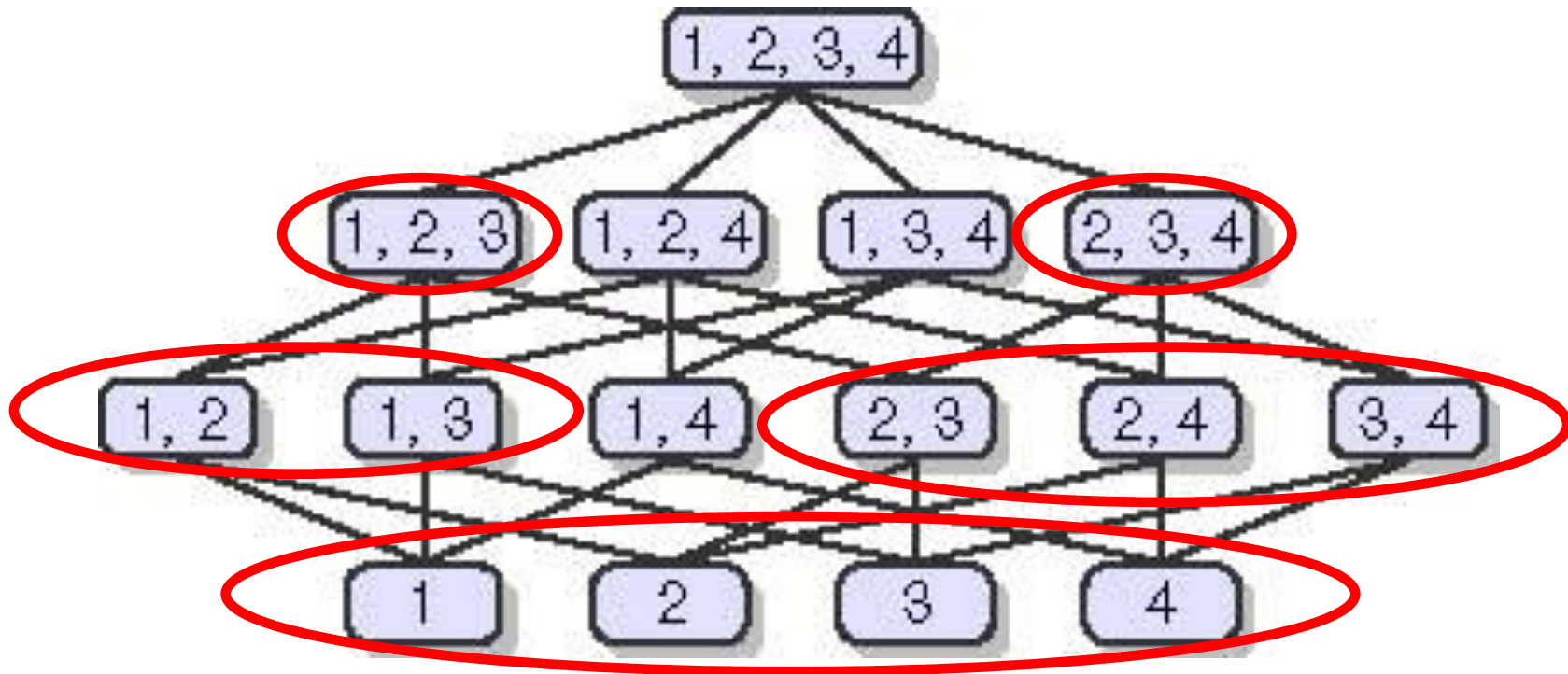
Score Discrepancy – a Glimpse at the Solution

- $\text{discrepancy}(O_C^+, O_C^-) \equiv \text{p-value of statistical test for } E[O_C^+] > E[O_C^-]$
- Different statistical tests exist.

RefOut – Summary

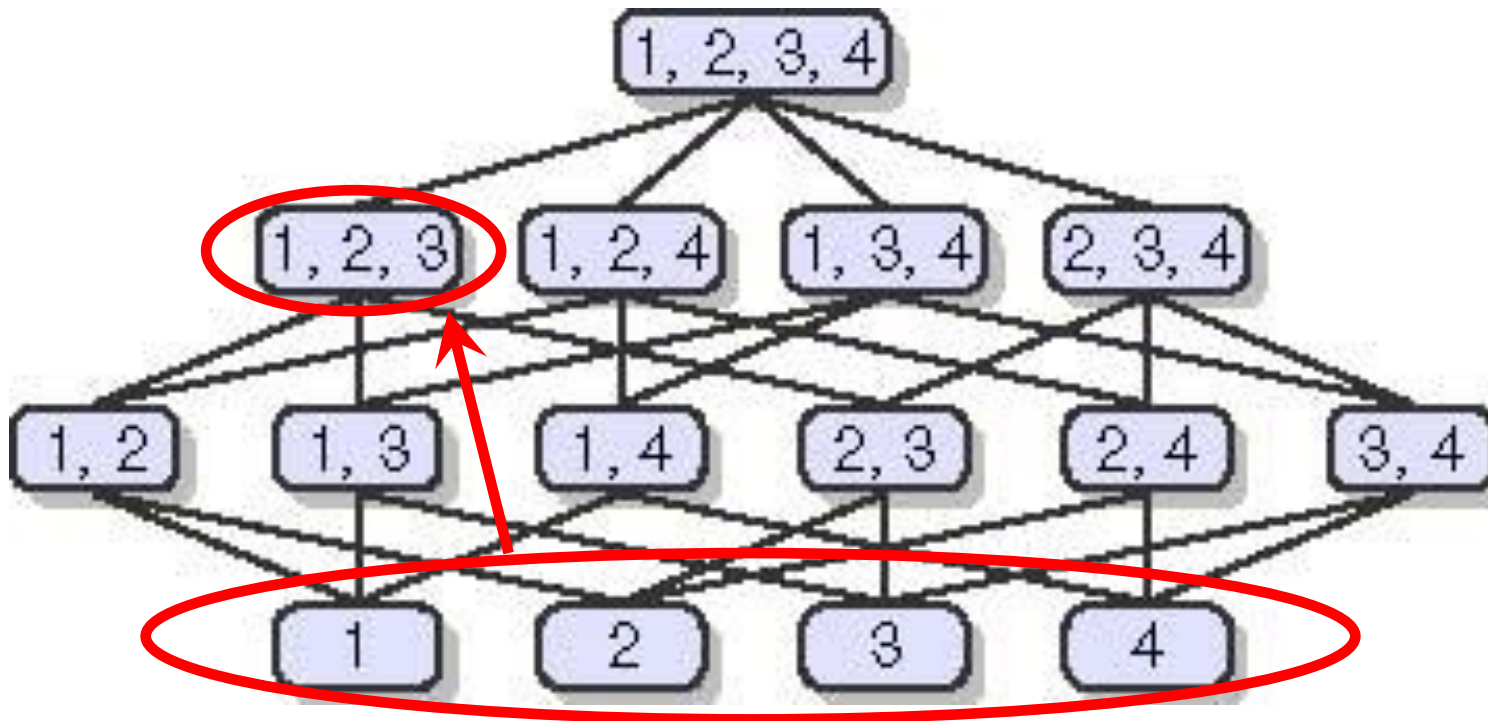
- Given an object, we seek subspace S where its outlierness/its score is maximal.
- We compute different subspaces for different objects. (Unlike HiCS.)
- *RefOut is adaptive* – scores of object in each subspace depend on the outlier model used.

Weaknesses of Existing Approaches (1)



- Heuristic illustrated here: APR (Apriori),
- HiCS makes use of APR.

Weaknesses of Existing Approaches (2)



Weaknesses of Existing Approaches (3)

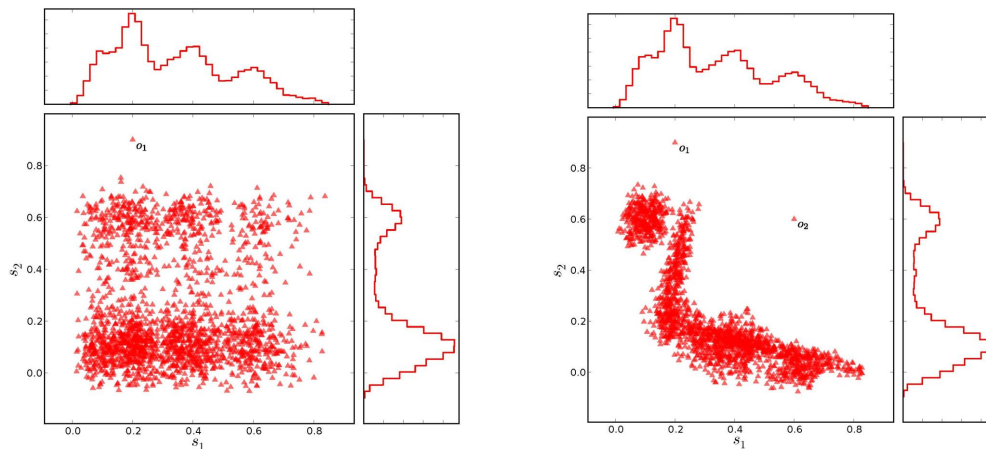
- Heuristic which subspaces are examined (e.g., APR (Apriori) used in HiCS)
 - effect on quality is unclear.

4S: Correlation of Two Dimensions

- Underlying notion: Correlation of Dimensions X and Y is:

$$\text{Corr}(X, Y) = \int_{-v}^v \int_{-v}^v (F_{XY}(x, y) - F_X(x) \cdot F_Y(y))^2 \cdot dx \cdot dy$$

F(...) is the cumulative distribution function (CDF).



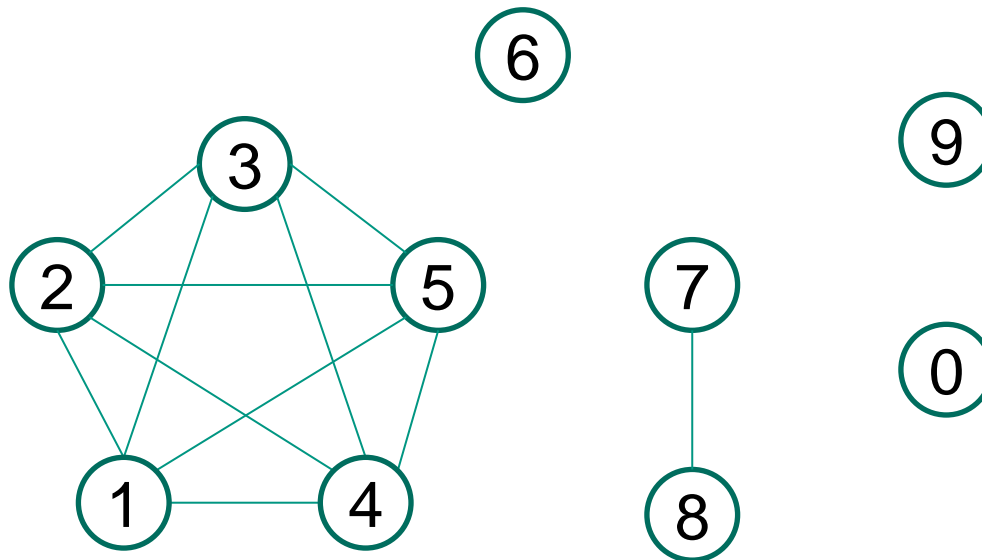
- As Corr is based on CDFs, it can be computed/estimated based on samples.

Correlation of Several Dimensions

- New definition of ‚correlation‘
for subspaces with arbitrarily many elements:
Subspace S has a high correlation if all pairs of dimensions from S
have a high correlation (a high Corr value).

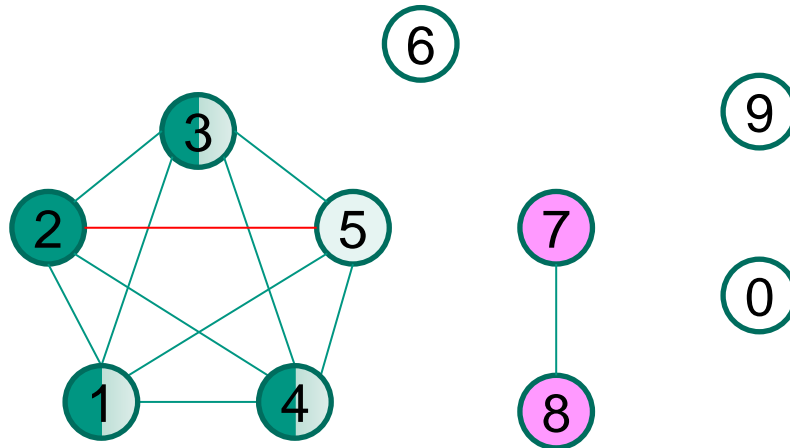
4S: Correlation Graph

- New definition of ‚correlation‘
 for subspaces with arbitrarily many elements:
 Subspace S has a high correlation if all pairs of dimensions from S
 have a high correlation (a high Corr value).



In other words, we look for cliques in correlation graph!

Consolidation of Results



- Maximal cliques frequently are only subspaces of interesting subspaces.
- Merge of ,similar' cliques/subspace. Column-wise clustering.

0	0	0	0	0	1
1	1	0	0	0	0
1	0	0	0	0	0
1	1	0	0	0	0
1	1	0	0	0	0
0	1	0	0	0	0
0	0	1	0	0	0
0	0	0	1	0	0
0	0	0	1	0	0
0	0	0	0	1	0

Efficiency Considerations

- APR can use Corr.
It will compute Corr of all pairs of dimensions.
- We are not interested in Corr of all such pairs, but only the K pairs with the largest Corr.
- To compute these pairs efficiently, we have proposed several useful bounds.

Fast Computation of Correlated Subspaces with k Dimensions

- APR with Corr yields subset of subspaces that fulfill our new definition. (Easy to prove.)
- I.e., approach based on our new definition does not yield results that are worse than existing ones.
- We have proposed fast computation of approximate result with small (guaranteed) loss of quality.

4S: Discussion

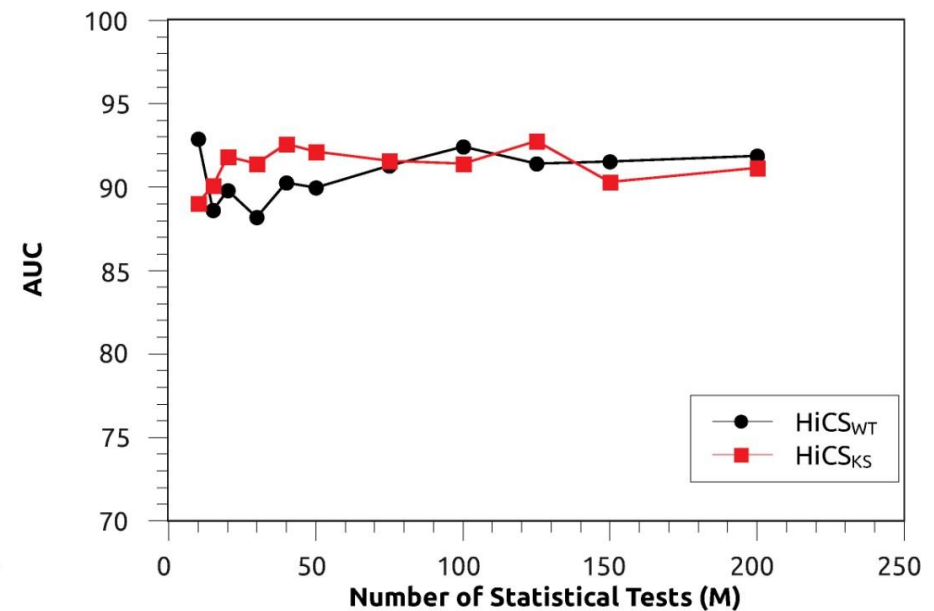
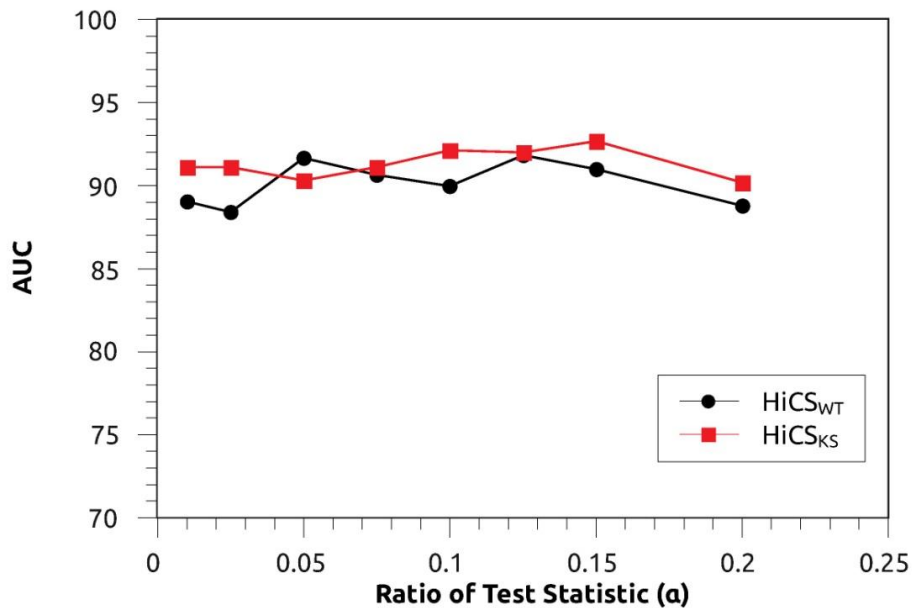
- Worst case complexity: $O(l \cdot (D - \log N)^3 \cdot N)$.
- Only worst-case, runtimes according to experiments are linear with N and sublinear with dimensionality.
- Addresses “the Challenge of Big Data in Science – with a Focus on Big Data Analytics” well.

Evaluation – HiCS

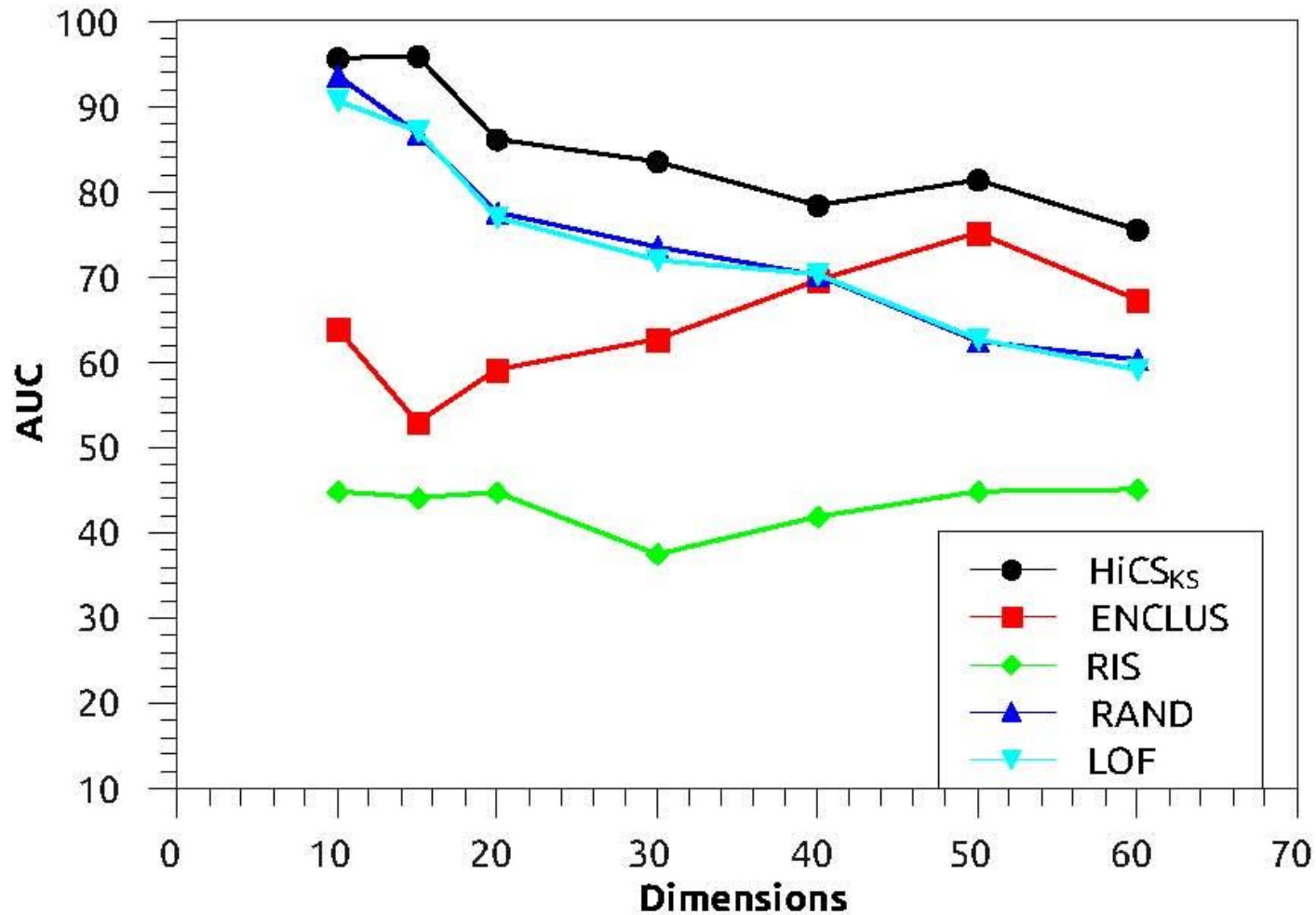
- Measure considered here:
Outlier-detection quality in step that follows.
Measured with Area under Curve (AUC).
- Method: Local Outlier Factor (LOF) with fixed parameterization.
- Subspace search:
 - none
 - Enclus
 - RIS
 - HiCS

Parameters

- HiCS is robust regarding parameterization in our experiments.



Effect of Dimensionality



Experiments with Real-World Data (1)

- Eight data sets frequently used.
- Result quality and runtimes.

Experiment	Runtime [sec.]				
	LOF	HiCS	Enclus	RIS	RANDSUB
Ann-Thyroid	7.1	37.2	68.1	574.0	674.0
Arrhythmia	0.5	26.4	7.9	2216.1	48.2
Breast	0.1	2.4	1.5	-	3.5
Breast (diagnostic)	0.3	15.8	11.8	14.3	28.2
Diabetes	0.3	3.3	5.9	4.0	26.2
Glass	0.0	0.2	0.3	0.1	1.7
Ionosphere	0.1	6.1	4.2	668.2	11.0
Pendigits	34.1	1194.5	2195.6	11282.7	3326.2

Experiments with Real-World Data (2)

Experiment	AUC [%]				
	LOF	HiCS	Enclus	RIS	RANDSUB
Ann-Thyroid	86.16	95.11	94.32	95.16	93.32
Arrhythmia	62.92	62.29	62.11	63.61	63.52
Breast	56.42	59.31	59.55	-	56.98
Breast (diagnostic)	86.94	94.23	94.19	90.77	87.07
Diabetes	70.98	72.47	71.15	71.63	71.70
Glass	76.86	80.05	79.73	80.65	78.48
Ionosphere	77.97	82.34	82.37	80.93	79.02
Pendigits	93.54	95.04	94.29	90.74	93.22

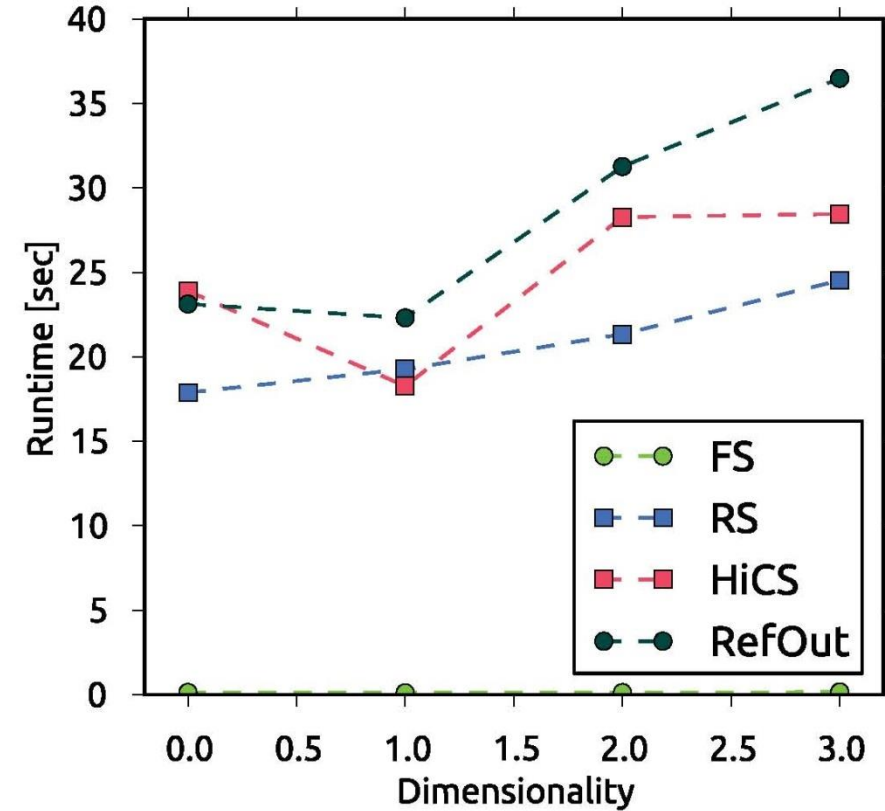
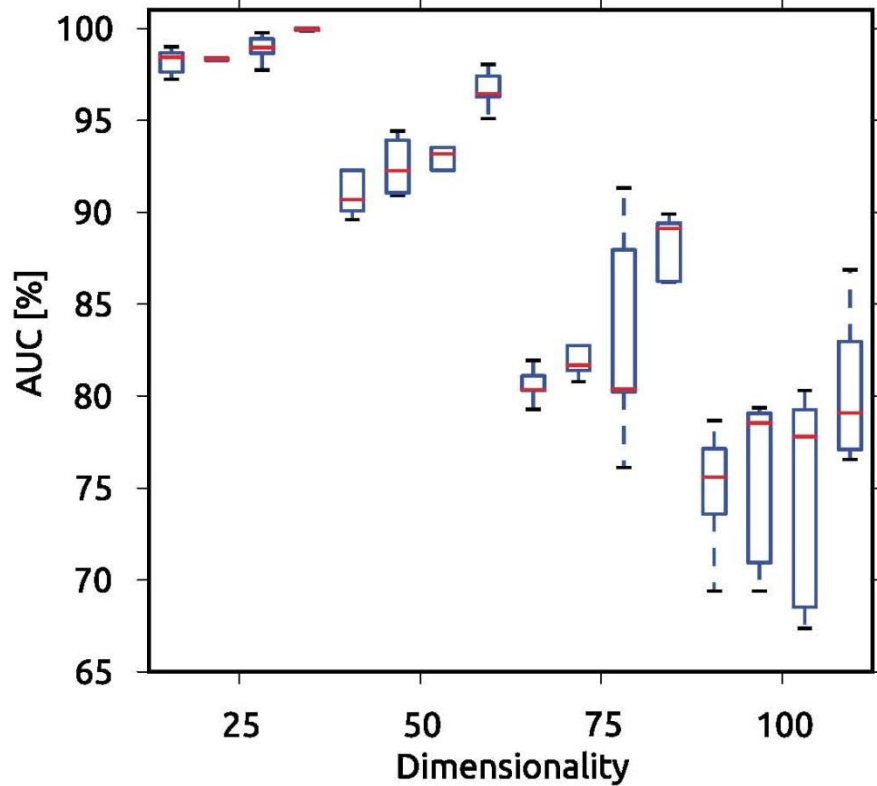
Experiments – RefOut (1)

- Measure considered here as well:
Outlier-detection quality in step that follows.
- Different outlier models:
 - LOF, density-based,
 - DB, distance-based,
 - FastABOD, angle-based.
- Approaches compared:
 - Full attribute space
 - Random subspace selection
 - HiCS
 - RefOut.

Experiments – RefOut (2)

- Both experiments with synthetic data and with real data.
- Ground truth with real data:
result of a – very time-consuming – exhaustive search of all subspaces with dimensionality $<$ threshold value.

Experiments – RefOut (3)



Conclusions and Outlook

- Useful results (domain-independent, broadly applicable) in the field of data mining.
- Our work complements domain-specific scientific work at KIT and elsewhere well.
 - In new scientific disciplines, support designing new models.
 - In established disciplines,
 - help identifying shortcomings of existing models systematically,
 - use our techniques to monitor complex experiments.
- In the future, we wish to continue/extend cooperations with scientists with big data.

Acknowledgements and Publications

- Fabian Keller, Emmanuel Müller, and Klemens Böhm:
HiCS: High Contrast Subspaces for Density-Based Outlier Ranking.
IEEE 28th International Conference on Data Engineering (ICDE 2012).
- Fabian Keller, Emmanuel Müller, Andreas Wixler, and Klemens Böhm:
Flexible and Adaptive Subspace Search for Outlier Analysis.
22nd ACM International Conference
on Information and Knowledge Management (CIKM 2013).
- Hoang Vu Nguyen, Emmanuel Müller, and Klemens Böhm:
4S: Scalable Subspace Search Scheme
Overcoming Traditional Apriori Processing.
IEEE International Conference on Big Data (BigData 2013).

Backup

Illustration – Discussion

- Outliers appear as such only in some subspaces.
- Some subspaces do not contain any outliers.
- True outliers occur in subspaces of different dimensionality.

⇒ Quality measure for subspaces sought.

I.e., how promising is it to search the subspace for outliers?

⇒ How to find such relevant subspaces in the first place?

Relationship to Big Data

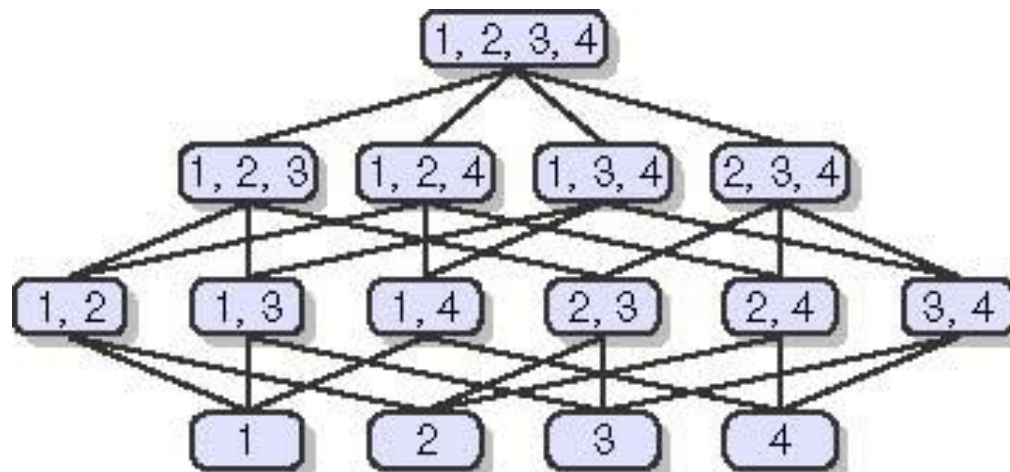
- Wikipedia: „Data sets so large and complex that it becomes difficult to process using on-hand ... tools ...“
- I.e., contingency on nature of problem, time available for analysis etc.
- Outlier detection is big data problem.

Comparison to Other Approaches

	Clustering	Outlier Mining
Fixed Space (Full Space)	<ul style="list-style-type: none"> • DBSCAN [KDD 1996] • ... 	<ul style="list-style-type: none"> • LOF [SIGMOD 2000] • LOCI [ICDE 2003] • ...
Coupled Method	<ul style="list-style-type: none"> • CLIQUE [SIGMOD 1998] • ... 	<ul style="list-style-type: none"> • OUTRES [ICDE 2011] • ...
Decoupled Method	<ul style="list-style-type: none"> • Enclus [KDD 1995] • RIS [PKDD 2003] 	<ul style="list-style-type: none"> • RandSubs [KDD 2005]

Challenges

1. Dimensional curse.
 - Many attributes.
 - Contrast is lost.
 - Outlier detection does not work any more.
2. Subspace Search
 - Exponentially many subspaces $P(A)$
 - Selecting relevant subspaces $RS \subset P(A)$



HiCS: Principle (4)

- More general than search for correlations:
Not only
 - linear
 - pairwiserelationships.

Contrast of Subspaces (1)

- For subspaces with little contrast, the following holds:

$$\underbrace{p_{s_1|s_2,\dots,s_d}(x_{s_1} | x_{s_2}, \dots, x_{s_d})}_{p_{s_i|C_i}^{(c)}} = \frac{p_{s_1,\dots,s_d}(x_{s_1}, \dots, x_{s_d})}{p_{s_2,\dots,s_d}(x_{s_2}, \dots, x_{s_d})} = \underbrace{p_{s_1}(x_{s_1})}_{p_{s_i}^{(m)}}$$

Contrast of Subspaces (2)

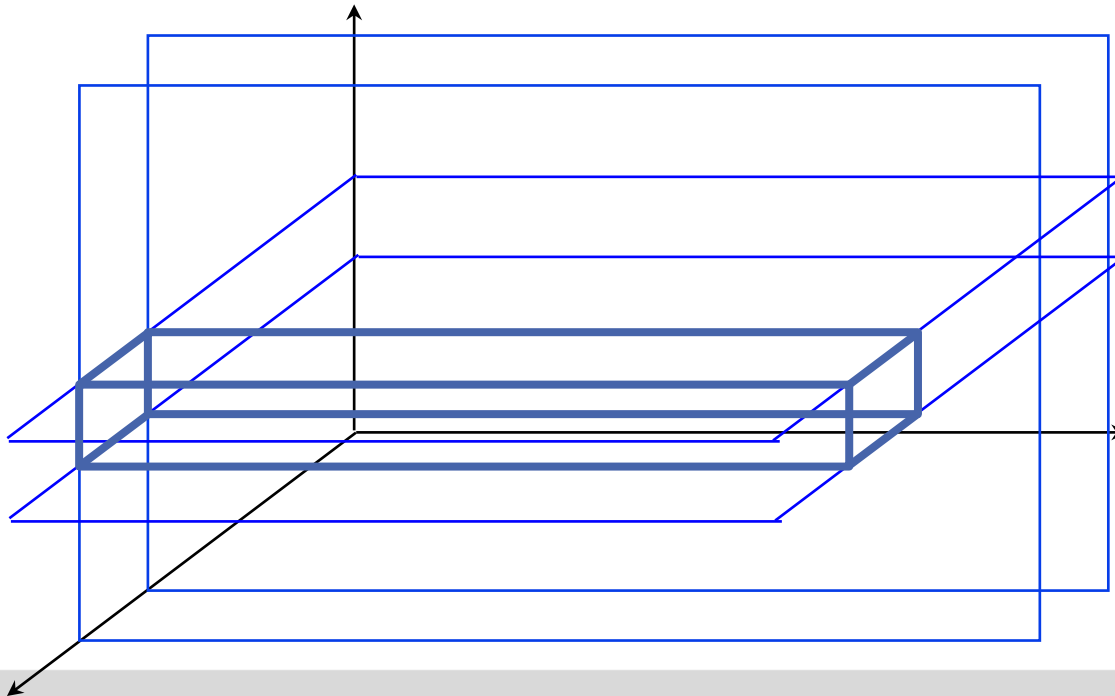
- Contrast – definition: Evaluate a subspace S using Monte Carlo algorithm with M iterations:
 - Choose isolated Attribute s_i randomly.
(s_i – horizontal axis in previous example.)
 - Generate set of conditions C_i randomly.
($\{, [0.1; 0.2]$ along s_2 ‘} in previous example.)
 - Check if equation on previous slide is violated.

$$\text{contrast}(S) \equiv \frac{1}{M} \sum_i^M \text{deviation}(p_{s_i}^{(m)}, p_{s_i|C_i}^{(c)})$$

- How to generate the C_i ?
- How to instantiate $\text{deviation}(p_{s_i}^{(m)}, p_{s_i|C_i}^{(c)})$?

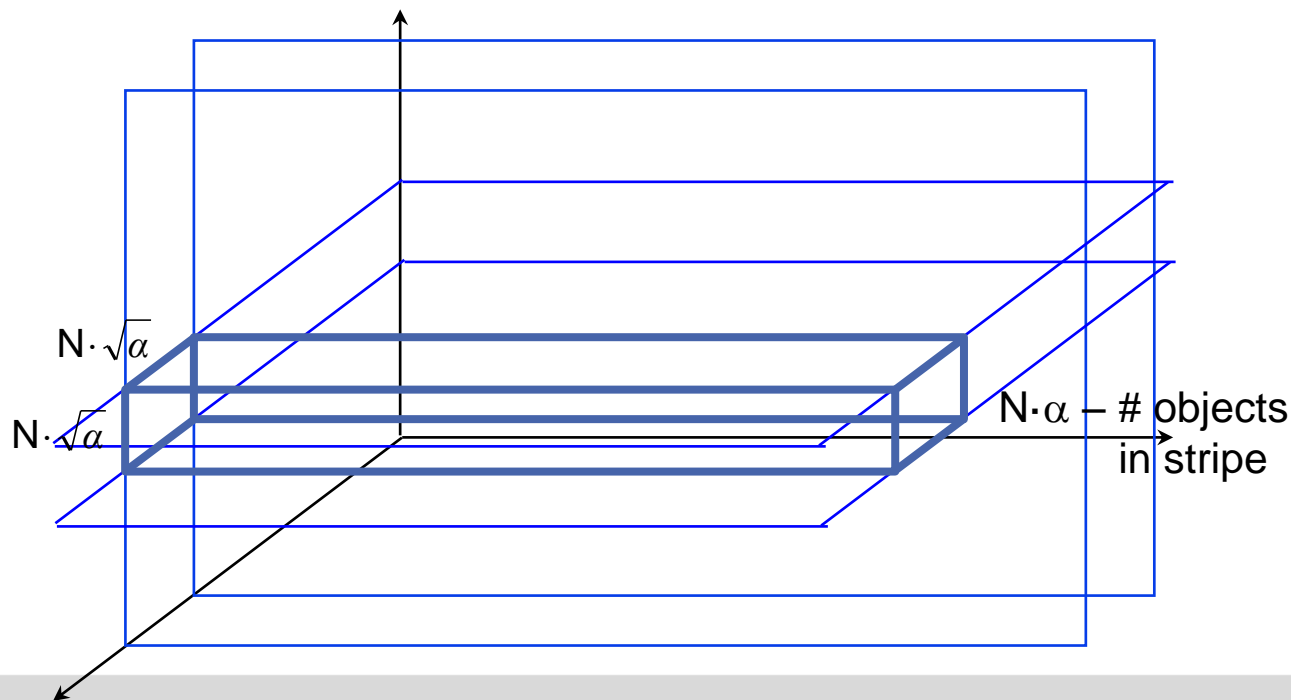
Adaptive Sets of Conditions (1)

- Let subspace S have d dimensions.
Then the number of conditions is as follows: $|C| = d - 1$
- Open issue: Set of conditions should adapt to dimensionality of the subspace.



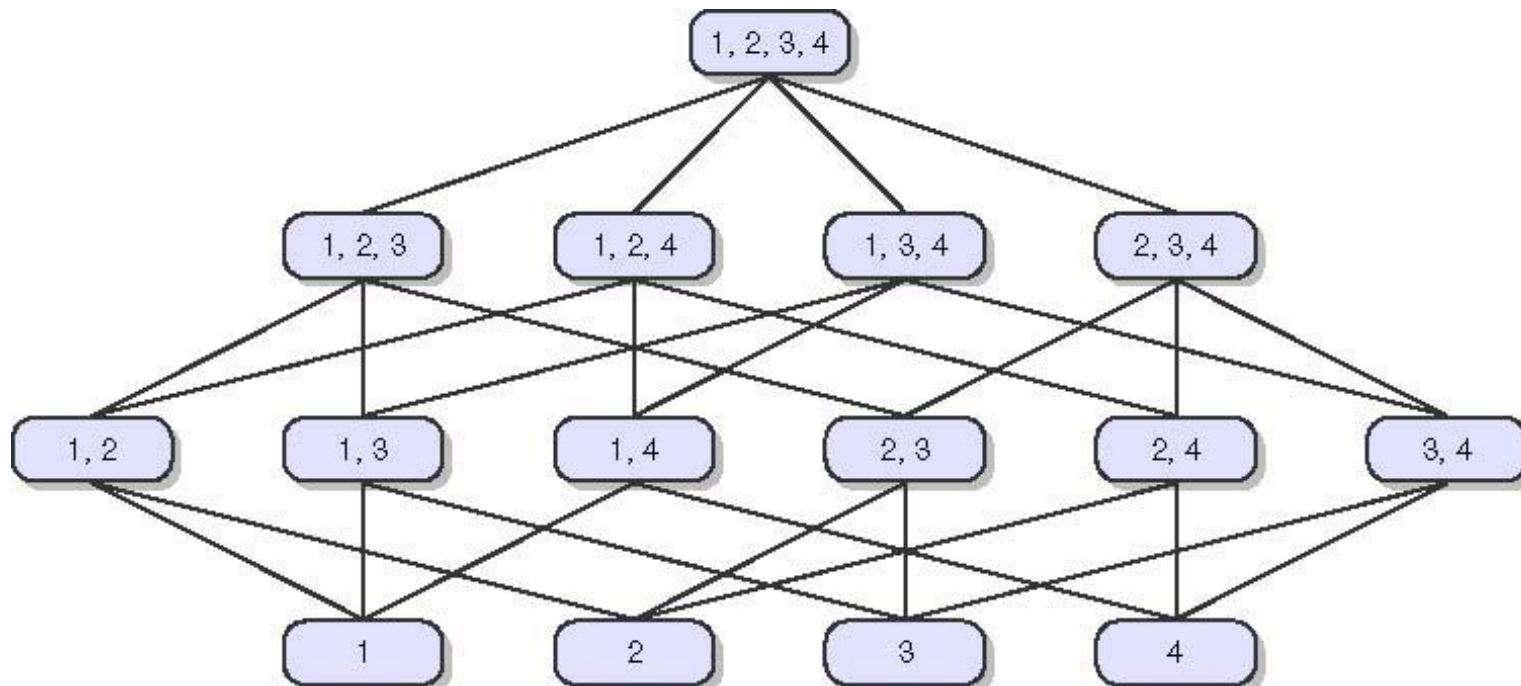
Adaptive Sets of Conditions (2)

- Our approach – ensure fixed sample size:
 - Choose position in the middle of the ‚stripe‘ randomly.
 - Choose width of stripe according to $N \cdot \sqrt{\alpha}$
 α is relative size of sample sought.
 - Width of stripes takes marginal distribution into account.



HiCS: Subspace Search

- Bottom up approach.
- Like Apriori, but without anti-monotonicity.



HiCS – Summary

- HiCS – a new method for subspace search for subspace-outlier detection.
- Subspaces are searched systematically, not randomly.
- Subspace is promising if distribution deviates from expected one, given distributions of its subspaces.

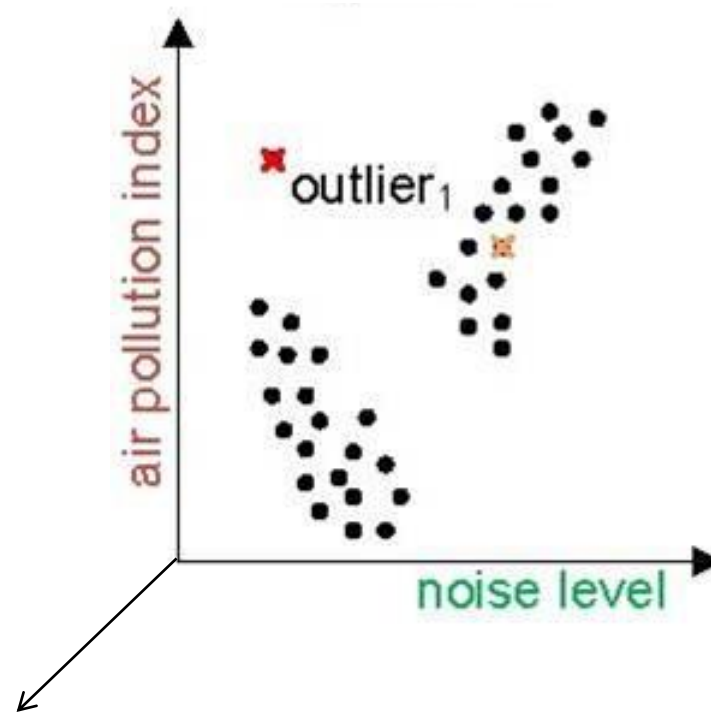
RefOut – Motivation (2)

- Respective experiment:
 - Taking all subspaces upto a certain dimensionality into account.
E.g., 206367 subspaces for ‚Breast‘
(31 dimensions, dimensionality of subspaces ≤ 5).
 - Compute outlierness of each object and for each subspace,
according to different outlier models.
 - For each object and each model,
we identify subspace where outlierness is maximum.
(Normalization, to ensure comparability between subspaces.)
 - The following table is abstraction of these results.

RefOut – Motivation (3)

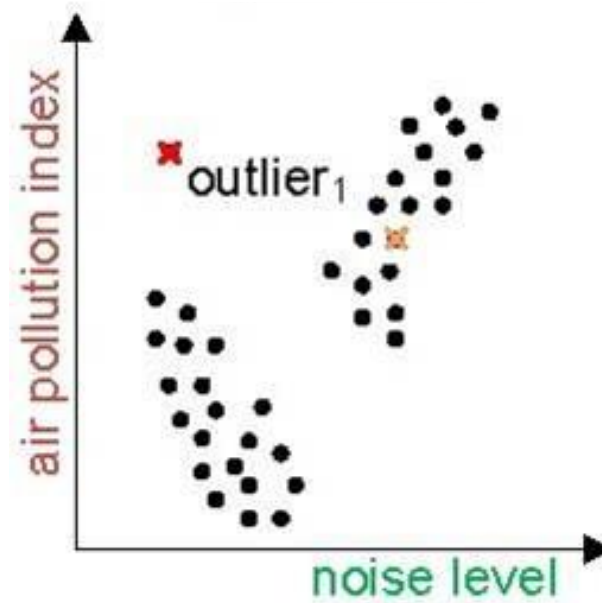
Dataset (size x dim)	Ground Truth	Peaks in Dim				
		1	2	3	4	5
Breast (198 x 31)	ABOD	0	139	40	16	3
	DB	58	81	44	15	0
	LOF	36	67	52	29	14
Breast Diagnostic (569 x 30)	ABOD	0	284	187	98	-
	DB	101	268	155	45	-
	LOF	94	177	177	121	-
Electricity Meter (1205 x 23)	ABOD	6	217	405	577	-
	DB	99	537	393	176	-
	LOF	197	374	413	221	-

RefOut – Idea (1)

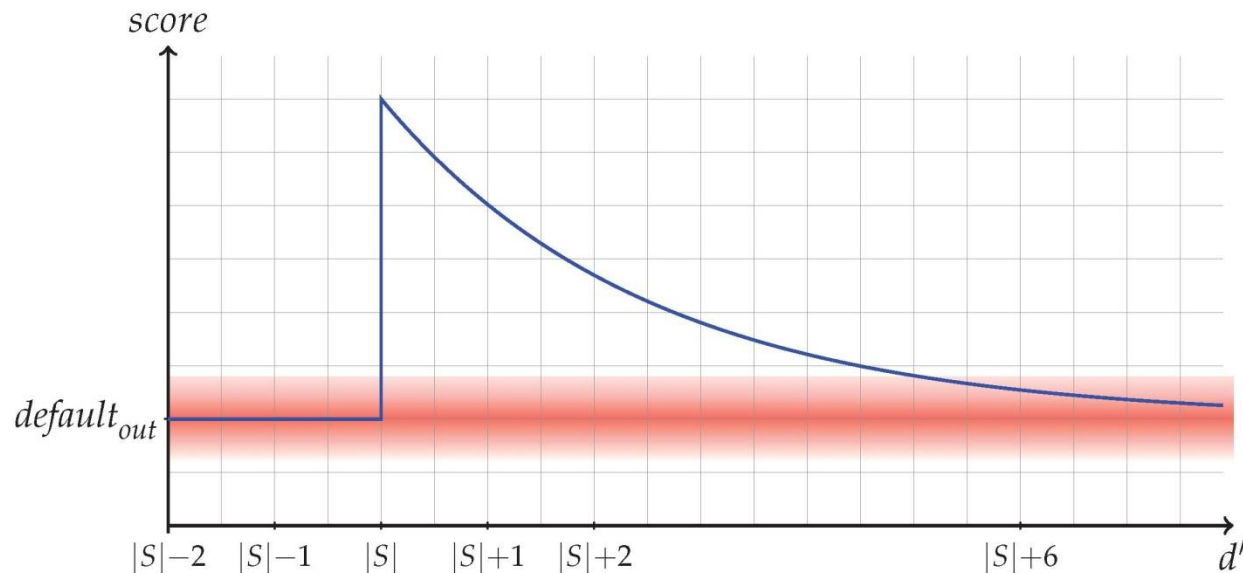


- One does not have to search unsystematically.
 If subspace contains outlier,
 a subspace of it could contain this outlier more prominently.

RefOut – Idea (2)



Characteristics of Outliers in Subspaces



- S – subspace where current object is prominent outlier.
- $T \supset S$: „dimensional curse“.
Scores decrease monotonically.
- Situation in reality:
 - One random subspace T per dimensionality d' at a time.
 - Different data sets and outlier models at a time.

Refinement – Continuation (2)

- There are $\binom{D}{d'}$ d' -dimensional subspaces possible (i.e., potential S').
- This calls for a heuristic.
 - Beam search with parameter *beamSize*.
 - $(d+1)$ -dimensional candidate
iff all d -dimensional subsets are in the result of the previous step.

Refinement – Illustration (2)

■ $O_{\{2\}}^+, E[O_{\{2\}}^+]$

Rank	Occurrence of attributes 1-12	Outlierness
1	[12 cells: 10 olive, 2 orange]	[brown bar]
2	[12 cells: 10 olive, 2 orange]	[brown bar]
3	[12 cells: 10 olive, 2 orange]	[brown bar]
4	[12 cells: 10 olive, 2 orange]	[brown bar]
5	[12 cells: 10 olive, 2 orange]	[brown bar]
6	[12 cells: 2 red, 10 green]	[blue bar]
7	[12 cells: 10 olive, 2 orange]	[brown bar]
8	[12 cells: 10 olive, 2 orange]	[brown bar]
9	[12 cells: 10 olive, 2 orange]	[brown bar]
10	[12 cells: 10 olive, 2 orange]	[brown bar]
11	[12 cells: 10 olive, 2 orange]	[brown bar]
12	[12 cells: 10 olive, 2 orange]	[brown bar]
13	[12 cells: 10 olive, 2 orange]	[brown bar]
14	[12 cells: 10 olive, 2 orange]	[brown bar]
15	[12 cells: 10 olive, 2 orange]	[brown bar]
16	[12 cells: 2 red, 10 green]	[blue bar]
17	[12 cells: 10 olive, 2 orange]	[brown bar]
18	[12 cells: 2 red, 10 green]	[blue bar]
19	[12 cells: 10 olive, 2 orange]	[brown bar]
20	[12 cells: 10 olive, 2 orange]	[brown bar]

Refinement – Illustration (3)

■ $Q_{\{1,2\}}^+, E[Q_{\{1,2\}}^+]$

Rank	Occurrence of attributes 1-12	Outlierness
1	[O][O][O][O][O][O][O][O][O][O][O][O]	High
2	[O][O][O][O][O][O][O][O][O][O][O][O]	High
3	[O][O][O][O][O][O][O][O][O][O][O][O]	High
4	[O][O][O][O][O][O][O][O][O][O][O][O]	High
5	[R][G][G][G][G][G][G][G][R][G][G][G]	Low
6	[R][R][G][G][G][G][G][G][R][G][G][G]	Low
7	[O][O][O][O][O][O][O][O][O][O][O][O]	High
8	[R][G][G][G][G][G][G][G][R][G][G][G]	Low
9	[R][R][G][G][G][G][G][G][R][G][G][G]	Low
10	[R][G][G][G][G][G][G][G][R][G][G][G]	Low
11	[O][O][O][O][O][O][O][O][O][O][O][O]	High
12	[R][G][G][G][G][G][G][G][R][G][G][G]	Low
13	[O][O][O][O][O][O][O][O][O][O][O][O]	High
14	[O][O][O][O][O][O][O][O][O][O][O][O]	High
15	[R][G][R][G][G][G][G][G][R][G][G][G]	Low
16	[G][R][G][G][R][G][G][G][R][G][G][G]	Low
17	[O][O][O][O][O][O][O][O][O][O][O][O]	High
18	[R][R][G][G][G][G][G][G][R][G][G][G]	Low
19	[O][O][O][O][O][O][O][O][O][O][O][O]	High
20	[O][O][O][O][O][O][O][O][O][O][O][O]	High

What do we have achieved so far?

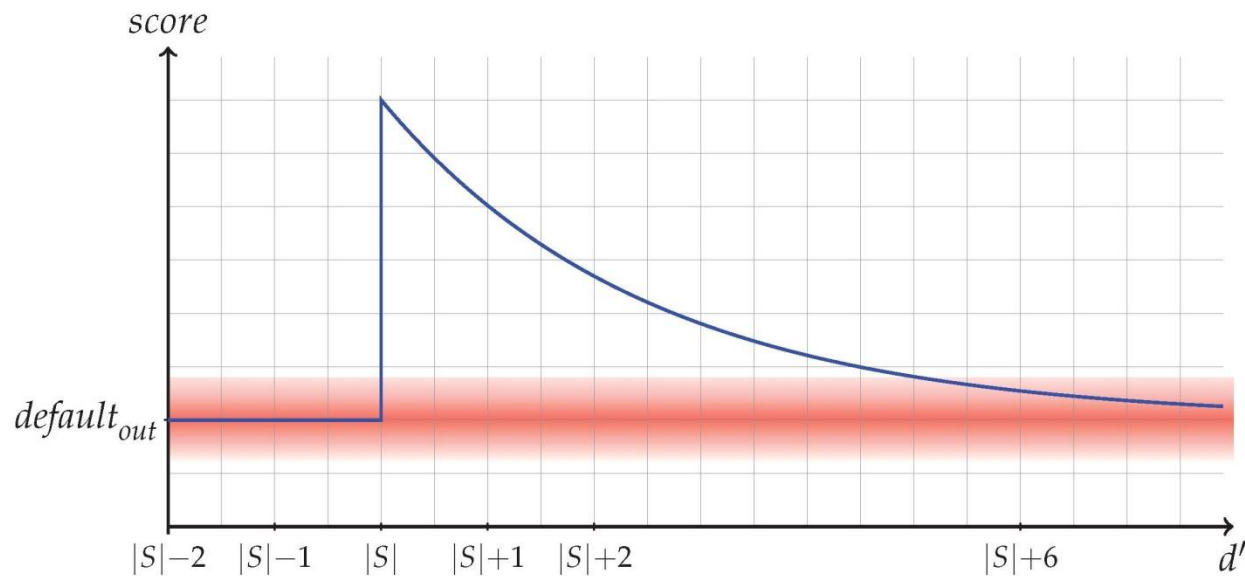
- x has been given.
- We know the score of x in each element of the subspace pool.
- Based on this, we can tell whether S or S' is more likely to contain x as a true outlier.

- However, we do not only want to choose between S and S' , we want to explicitly find a good S .
- Try different S systematically. See paper for details.

Rank	Occurrence of attributes 1-12	Outlierness
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		
17		
18		
19		
20		

Score Discrepancy Problem – Size of T

- $|T|$, dimensionality of elements of subspace pool, is small.
 T and S are similar, $\text{score}(\bar{x}_T) > \text{default}_{out}$



- On the other hand: Probability that $T \supset S$ is very small.
 To illustrate, let $S = \{1, 2, 3, 4\}$. If $|T| = 5$,
 with dimensions chosen randomly, it is unlikely that $S \subset T$.

Refinement – Eventual Candidate Selection

- Rank all subspaces inspected irrespective of their dimensionality by their discrepancy(...).
- The first d' dimensions of the subspaces with the highest ranks form the subspace sought.

RefOut – Summary

- Given an object, we seek subspace S where its score is maximal.
- Given a (randomly chosen) subspace pool, we try to partition it as follows:
 - One partition contains all superspaces of a subspace S' , the other one the other pool elements.
 - The current object has high scores in the elements of the first partition, but low ones in the elements of the second one.
- This lets us approximate S .
- Thus, we compute different subspaces for different objects. (Unlike HiCS.)
- RefOut is adaptive – scores of object in each subspace depend on the outlier model used.

4S: Underlying Concepts (1)

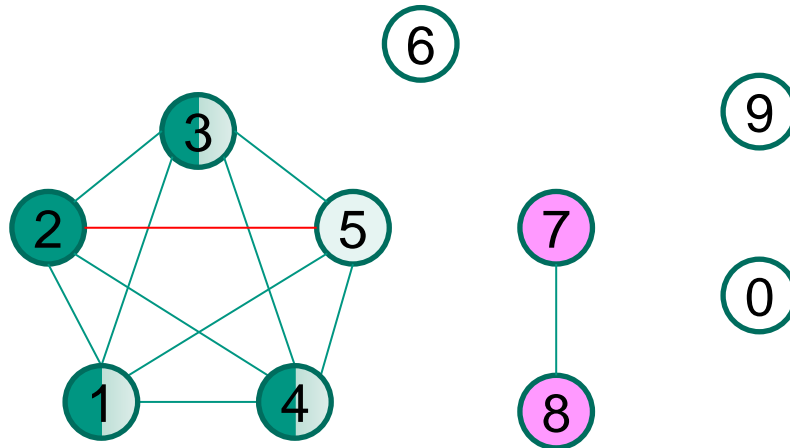
- Total correlation of $\{X_i\}_{i=1}^d$ is:

$$T(\{X_1, \dots, X_d\}) = \sum_{i=2}^d H(X_i) - H(X_i | X_1, \dots, X_{i-1})$$

H is Shannon (differential) entropy.

- Conventional notion.
- Requires estimating the probability density functions, by means of, say, discretization.

Consolidation of Results



- Maximal cliques frequently are only subspaces of interesting subspaces.
- Merge of 'similar' subspace. Column-wise clustering.
- Use MDL-based clustering technique: Effort to describe the columns, given the clustering result, is minimal.

0	0	0	0	0	1
1	1	0	0	0	0
1	0	0	0	0	0
1	1	0	0	0	0
1	1	0	0	0	0
0	1	0	0	0	0
0	0	1	0	0	0
0	0	0	1	0	0
0	0	0	1	0	0
0	0	0	0	1	0

Fast Computation of Correlated Subspaces with k Dimensions

- APR with Corr yields subset of subspaces that fulfill our new definition. (Easy to prove.)
- I.e., approach based on our new definition does not yield results that are worse than existing ones.
- We have proposed fast computation of approximate result with small (guaranteed) loss of quality.